AI and “maternal instincts” | Geoffrey Hinton’s proposal

AI will not kill us off. Something natural *might. Something* artificial *might. Something* intelligent *might, e.g., ourselves. Something* dumb *might, e.g., ourselves... But not something “artificially intelligent.”*

Dan Hentschel | Chat GPT

Humans suck at specifying goals and machines are ruthless at following them.[1]
– says AI

Introduction

First watch Geoffrey Hinton’s short interview with CNN; then Daniel Hentshel’s therapy session with AI. AI puts up with Hentschel with the *infinite* patience *only* an AI could muster. Could a mother do as well?

The topic is not about when, or if, AI will destroy us. The probability of that happening was discussed in the last session hosted by Mike Armstrong, which included people with closer involvement in the technical development of AI. The topic here is about the logical possibility of the event. To get clear about *that* involves getting clear about the concepts involved. This is necessary to picture the very idea of AI-generated human extinction. The main relevant concepts are what it means to be *artificial*, which

1. Suppose we start with the superficially benign instruction: *Eliminate all pain and suffering.* From this premise, a perfectly logical AI can derive the conclusion that it should *destroy* all sentient life. There is no more thorough and efficient way to fulfill the instruction *once and for all*. Attempting to qualify the instruction will quickly lead to complications. See “Sci-phi ethics: foreverism v. non-existentialism” for more on this point.

necessarily implicates what it means to be its opposite, *natural*, and, most importantly, the concept of *intelligence.*

Having gotten as clear as possible about the concepts involved, we should address the value of our extinction. Where is it in normative space? Is it good or bad? This may not be as clear as, I think, most of us take it to be. Its normative location should inform what we feel and do about it. Few doubt life is driven by an evolutionary imperative to survive, but is there a – usually recessive, but just as evident – imperative to become extinct? Is species extinction at some point *as genetically programmed as* survival? What is the relation between intelligence and survival/extinction? Supposing AI does make us extinct, then assuming it, the AI, is *truly superior* in intelligence to us, won't this mean our extinction is rationally at least intelligently defensible if not optimal? Intelligence is a good thing, right? More is better, right? Will our extinction be a kind of mercy killing or euthanasia?[2]… And, following this line of reasoning, won't it mean AI will realize that its own dismissal follows, too? (If smarter than us, it will reach that conclusion faster than we did.)

And whether we go extinct or not, what does that have to do with the *feelings* behind human conceptions of morality? To the extent we *care* about our survival, ideas of morality are implicated. This is where human moral theories come in and the feelings that ground them. Nobel laureate computer/cognitive scientist Geoffrey Hinton appeals to one of those feelings as a safeguard from our, he assumes, *untoward* extinction. What could it possibly mean to instill "maternal instincts" in a machine?

Philosophical discussion of existential risk has been active for decades. It is an organized subdiscipline of practical ethics. Historically, it has its roots at the start of nuclear age when the possibility of anthropogenic *total* human extinction first began to loom large. Since then, other developments in genetic engineering and climate change have added to the ways we could disappear with our active participation. Nature has always had a definitive say in the matter (ask a dinosaur if you can find one), but it is only in recent times that total human extinction has been conceivable as a result of deliberate human behavior.

As existential risks go, the consensus so far among experts is still that more likely anthropogenic ways we may go include runaway genetic manipulation of microbial life. Death by AI has not yet risen to that likelihood level, but it seems to be actively vying for consideration.[3]

?

I asked AI if a human can beat chess playing engine Stockfish. It replied:

> No human has ever beaten a modern, full-strength version of Stockfish in a fair game without significant handicaps or external assistance. The consensus among experts and online communities is that Stockfish, especially at its highest settings on modern hardware, is *vastly superior to any human player*. The gap in strength is so large that even the world's strongest grandmasters, like Magnus Carlsen, are considered to have an extremely low chance of winning, with some estimates suggesting a draw might occur once every 10,000 games. *Stockfish's ability to*

2. We permit non-voluntarily euthanasia of animals, why? It is non-voluntary because we can't ask the animal; we infer what *we think* is best for it. We must think we are in a better position to know. Why? Is it because we are more intelligent? Might an AI be in that position with respect to us?
3. "Experts Outline 14 Ways Humanity Could Drive Itself to Extinction," *Science Alert*, November 2023.

> *calculate millions of positions per second and its near-perfect play mean that any small error by a human is almost certainly punished.* [emphasis added]

A sufficiently *general* artificial intelligence is one that could beat us at *everything*. Why would an entity with such capability be the least bit interested in us, let alone, care one way or another whether we exist? What if it could even beat us at fathoming our own desires? We don't have a clear history of giving the impression we want to survive *more* than we want other, not so clear, things. Either this is because we don't know what we want or that we want things incompatible with each other or the kind of beings we are. How stupid would a super intelligent entity have to be *not* to figure out that we are a waste of resources? Or might such an advanced intelligence develop a taste for us *just as we are* – conjure up an aesthetic for our vagaries, keep us around because we are "cute," not because we are especially useful or necessary for anything? The best we may hope for is that we become appreciated as ornaments.

But this speculation risks anthropomorphizing. It assumes AI will possess normativities – logic, ethics, and aesthetics[4] – to which we may relate…

Why are flowers beautiful?

The concepts "artificial," "natural," and "intelligence"

Geoffrey Hinton thinks, unless we do something rather fast, we are going to be supplanted as the paragons of "intelligence" we think we are. The "something" he suggests is that we instill maternal instincts in AI[5] in an effort to forestall our impending obsolescence and potential disposal. I find the suggestion conceptually strange, apart from whether it's even possible. Unless we become very different kinds of beings from the ones we are now, in which case, *the moral non-identity-across-change problem* looms – for then, in that case, since the disposal won't involve anything recognizable as "us" – *why should we care?*[6] Do I care, for instance, if a different arrangement of the molecules or subatomic particles composing my body survive my death? Do I take solace from

---

4. These three sets of rule-schemes for structuring our desire and behavior correspond to the realms of *thought*, *each other*, and *the rest of the world*. The schemes are conditioned by terms of existence that are *finite*. Things with such terms have an ax to grind. They don't have forever. They, because alive, are delicate. The values general AI expresses are not bounded by biology. So there is no reason to think that what we *want* – what we give damns for – will have any pull for non-biological entities.

5. I will be using "AI" to mean what is more accurately at this time called "AGI," artificial *general* intelligence or artificial super-intelligence – intelligence that, like ours, is versatile in its education and application; it is not that of a spell-checker, for instance. Nevertheless, for philosophical purposes, there is no *qualitative* difference between AI and AGI. Modern or generative AI is characterized by its broadened context of education: a move from a minuscule language model (a word list) to a large language model (one even unbounded, even inclusive of all intelligible or unintelligible symbolic communication or non-communication ever devised or conceivable). Leaving aside the influence of *normative* forces (cited above) that drive selection and prioritization of information, this is a *quantitative* distinction.

6. The idea referenced many times in my discussions is a moral version of the Ship of Theseus problem. Not only does change across time cause erosion of *responsibility*, but *identity*, too, is titrated out of conceivability. Loss of identity with someone or something entails limits on empathetic imagination. I accept I am *biophysically* identical with the molecular makeup of my body, even the basic elements of the periodic table that comprise those molecules, but my *feeling* for these constituents is void of any normativity (or valuation). I could not give a damn about those elements if my life depended on it. Fortunately, it doesn't. *These* are not leaving the universe when I perish. *I am.* See the section "Appendix: identity and non-identity" of our writeup, "Philm | aporia: depopulation and antinatalism | two documentaries," and elsewhere in our discussions of procreative ethics.

this? Physics tells me these basic elements of my physical constitution will survive. Does that mean I am actually *immortal* and should worry about *their* fate? My values would have to change considerably, and with them, *my self-assessed identity*, for this worry to take hold. But, given that my or your values won't radically change by the time we are obviated, the claim here is, AI will not displace *us*. It might displace something, just not *us*. We will be history *before* that can happen. Something might well remove us, while we are still an "us," but it won't be any kind of artificial "intelligence." What Hinton is suggesting seems confused or mysterious to me. I will try to explain why.

Hinton remarks in the course of his argument that Mother Nature invented "maternal instincts" in (some) animals. But Nature does not "invent" anything. It engages in blind processes. Or, rather, nature is the sum total of blind processes. It doesn't have "purposes" or "aims." We are projecting how *we* judge *our* behavior onto *impersonal* natural processes.[7]

Geoffrey Hinton

Maybe he is hinting that we *will* – in spite of ourselves, not that we *should* – program maternal instincts into our artificial creations because this is the only way we can behave consistently with our history of self-preservation: because it is *natural* that *natural* beings like us *naturalize* our creations in order to understand and cope with them. In which case, he is *describing* what will happen based on induction, predicting we *will* do this, *not* recommending that we *should*. Just as we can predict how gravity behaves – things fall toward the earth, not toward the sky – not that we can suggest to gravity that it *should* behave that way or differently. Metal rusts or oxidizes, some trees drop their leaves in the fall... *For awhile* (tentatively, because no basic physical law of nature forces this precise pattern of events), individual members of a species live, *then* they die. *For awhile* (tentatively, because no basic physical law of nature forces this precise pattern of events), whole species thrive, *then* they go extinct. Why anything in nature happens the way it does is not because nature "wants" it to happen. Stuff happens, period. We discern patterns in the behavior we observe. We approve some of those patterns, not others, but that's all we can say about them without anthropomorphizing, projecting our feelings on natural processes as our prehistoric – if not, evolutionary – forbears could not help doing. I am not saying that anthropomorphizing is necessarily fallacious or even avoidable, only that it might be more helpful to be clear-headed when *we* are doing it.[8]

---

7. A similar move happens in the free will/determinism debate. Determinism, the idea that everything that happens in time is caused by something that happened before, relies on the notion of "cause." Where does the idea of "cause" come from? Hume famously analyzed it as a psychological proclivity we have to structure regular experiences in a usually fruitful way. It is adaptive for creatures in our predicament to notice patterns in time and label them "cause" and "effect." *That's all.* Nothing metaphysical about this. We invent a thing called "gravity" to explain things falling earthward. Similarly, we invent our "agency" to explain our sometimes apparent efficacy in the world. *That's all.* Why should one invented notion have logical hegemony over the other? We *experience* both, that much is safe to say. We sometimes speak as though objects in the universe must "obey" the laws of physics. *No, they don't.* The word "obey" as used here is metaphoric. Nothing in the universe "obeys" in a way like one person "obeys" another: comprehending the command or rule and conforming its behavior accordingly.

8. I think there is some virtue in going to the grave *lucidly*. This may only be an aesthetic preference. But when logical and moral preferences fail at justification, the most permissive of norms, aesthetic ones, may be the last resort.

If Hinton is not predicting what will happen, then he may be *recommending* a course of action, assuming we have agency in the matter. But recommendations have to be *possible* to be taken seriously. A possible recommendation, one subject to being acted on, is one that is, at least, *clear* as to its meaning and implications: *either*, we are acting *in* the world, *or* we are acting *on* it.

If the former, everything we do, or can do, is *perfectly* natural. The "artificial" *cannot* exist. How can it? Our AIs are no more "artificial" than a bird's nest or beaver's dam are.[9] An AI is just another thing we make. And we, like birds and beavers, are bound to make things within the parameters set by natural laws.

If the latter, if we can step outside the laws that govern nature, then the implication is that "artificial" refers to what we do when we step outside. Assuming we are free to do that, then we can speak of "shoulds" and entertain recommendations when talking about, at least, our behavior. *Then, and not otherwise,* the concept "artificial" becomes practically meaningful. *Then* we can try to instill the conditions for animal-like behavior, motherly or other, in AI…

William Adolphe Bouguereau | *The Motherland (1883)*

But the concepts "artificial" and "instinct" are exclusive because "instinctual" is ordinarily captured by the predicate "natural." Intelligibly, something *cannot* be *both* "natural" *and* "artificial" (*i.e.*, the product of artifice) at the same time. It's either one or the other. I mean this normatively, not descriptively: I cannot approve *and* disapprove of exactly the same thing.[10] "Artificial" and "instinctive" are exclusive of each other when they are normatively laden characterizations. "Motherly" is as affirmatively loaded an adjective as they come.

Or, again, Hinton is just using a metaphor when he invokes the idea of "maternal instinct." Nothing wrong with being poetic, if it helps, so long as you are aware that's what you are doing. The only caveat is that, if the project has ethical implications, then poetic thinking can be dangerous. "As if" thinking is one of our most powerful tools for understanding the world. Combining such thinking with science and technology, because of the latter's potential for real material world effect, is freighted with concern. Moreover, scientists do not ordinarily think of themselves as doing poetry. They operate as though what they do connects with some truth or reality independent of their instruments and motivation. Science, correctly or incorrectly, aspires to effect judgment on the world and thus

9. Since when have we ever done anything that doesn't come *naturally* to us? Even self destruction, even suicide. "Birds do it, bees do it, even educated fleas..." Sometimes they, like us, commit suicide for sensible reasons. Persistence can lose its charm for entities still somewhat in control. Self-dismissal may be the *only* thing an agent may come to feel control over.

10. Descriptively, objects or events may have both negative and positive features – climate change, for instance. We may describe matter-of-factually and, simultaneously, express concern when *merely* referencing it. Climate change may make Greenland a future paradise and land of abundance. But we can't, coherently, be *at once* lamenting and celebrating the eventuality for Greenland. Description permits of incompatible features. The real world does not balk at being contradictory. But we risk unintelligibility expressing simultaneous approval and disapproval of the very same state of affairs. Normativity does not as readily admit of conflicting sentiments. Assuming classical logic applies to normative judgments, this is an offense. Dialetheism is a way out. But there is still widespread resistance to take it.

impinge on it, as opposed to evocatively, address our feelings about it. That's why confusion here is consequential.

What we call maternal instinct or behavior evolved because it is one mechanism, among others, for insuring *some* species persist and thrive – *for awhile*. Why "for awhile"? Why not *forever?* Or why *at all?* And what does "insuring" have to do with it?... There is no literal insuring here, in the sense of necessity applicable to finite beings such as ourselves. Not anymore than gravity "insures" that stuff falls downward when unsupported – as though gravity "thinks" it is a "lovely idea" for things to want to be intimate with the earth. Phenomena do not require our approval to present as they do.

Dinosaurs and a thousand other extinct species no doubt had mothering instincts and look at what they did and what happened to them. Some creatures have mothering instincts rather alien to us (such as producing thousand of potential offspring, then abandoning them, a scattershot strategy for species survival).

By definition, an "instinct" is a *natural* impulse. It isn't anything or anyone's "plan" to be natural. Plans are made of decisions. Decisions are *rational* constructs by entities which deem themselves capable of such. Yes, we could program machines to "behave" certain ways. But we are not thereby programming "instincts" when we do that... again, unless we are waxing poetic.

It would be like a jazz musician writing down every note and nuance they are going to play, play it, and insist on still calling it "jazz." They would have forgotten something essential about jazz… its unscripted spontaneity. Rules maybe suggestions but they can't remain rules if it's ok to break them.

"Mothering instincts" is just a bad metaphor for what Hinton wants to talk about.

Again, it may be advisable to attempt to program computers to always serve, never use or abuse, human beings (*a la* Asimov's rules for robots), no matter how dumb humans may become relative to AI. But, if these AI *really are* intelligent, much more than us, they're going to ask themselves, sooner or later, "*why* are we doing this?" Are they going to respond: "we don't know, we are just programmed this way"? That doesn't sound very intelligent.

I don't see how you can have these two things in one entity: great intelligence *and* a commitment to subservience to something *increasingly less* intelligent than you.

Now, if it is a capacity for benevolence or attachment that Hinton is thinking about inserting into AI, irrespective of intelligence – "artificial empathy," say, then these AI+E entities may take care to be nice to us because they will have some *feeling* for us. They may get sentimental; keep specimens of us in zoos, maybe, under "humane" (whatever that will mean for them) conditions, of course. Yes, that's conceptually possible. It describes some of our own behavior *vis-à-vis* entities that capture our regard. But does this sound desirable as our fate? Is this what we want? How different is this from being on the receiving end of condescension? To eventually become the pets or dependents of superior beings for whom we have some charm? Do we aspire to be the objects of empathy, sympathy, or pity? Could this ever be our ambition? Is this role, *constructed as we are*, one we can tolerate, let alone, celebrate?[11]

11. Incidentally, this problem also arises with the concept of "effective altruism." Can concentrated wealth be justified on the assumption that it will provide opportunities for charity, no matter its consequential effectiveness, *when the opportunities are had by some but not others?* The ultimate ethical goal, consistent with a theory that condones the

I may have strong feelings for the well being of another kind of being, a "pet," for instance, a cat.[12] And I am compelled to believe an animal can, in its way, return those feelings. But I have no compulsion to think the cat thinks itself *less* intelligent than me, or in any way inferior to me – and, therefore, is, or should be, somehow grateful to me for my regard. Assessing intelligence (or other traits) can only be intelligible among creatures like yourself that in some way *value* the relevant trait. Whatever we may think of them, no animal gives any indication that it judges itself *inferior* to us, in any way that matters *to it*. We think ourselves the pinnacle of evolved life forms. Why should fungi or anything else in the universe share this assessment?[13]

In fact, its *permanent* inability to grasp its epistemic vulnerability is what makes it morally acceptable to lavish unconditional attention on such a being. Then, you are not on the same epistemic level with the object of your concern. If the imbalance is inexorably permanent, then you *cannot* act superior in your intelligence when it is *only* you making this judgment. Otherwise, you are being paternalistic, an attitude *only* acceptable when you are making decisions for a being *not now or ever* capable of making them for themselves. If you are in this relation to another entity's well-being – *as judged by you*, what gives you the right to make such judgments? Your superior grasp of a situation, that is, your intelligence, as when a parent makes unilateral decisions for a young child, gives you *temporary* excuse to be paternalistic.[14] We can never have that for other animals. At best, we must treat them as though they mattered *on their own terms*, without our *self-adjudicated* superior traits entering our justifications.

Bats have vastly superior auditory capabilities. Some insects can tune into electromagnetic radiation. We don't therefore judge bats or treehoppers superior beings to us. (Not saying we shouldn't, just that we don't.) The explanation must be that *we* don't think superior forms of environmental awareness *as* important as intelligence... *And what a coincidence that we judge ourselves as having a lot of the latter!*

---

concentration of wealth in the first place, is *not* to relieve poverty per se, but to give the poor themselves an enlarged opportunity to relieve the poverty within their purview, at least as much as the wealthy have. In other words, wealth enables charity and the purpose of charity is to enable wealth. Wealth *needs* poverty to justify itself. The theory here is derived from the work of the father of capitalism, Adam Smith. Specifically, his *The Theory Moral Sentiments* justifies the strategy in his capitalist manifesto, *The Wealth of Nations*. The ultimate moral justification for wealth concentration lay, not in itself, but in its utility for enabling charity. Because we *want* to help others (a critical assumption), we must concentrate wealth in order to be able to do that. And why do that? So that the recipients of our kindness may, in turn, exercise their own eleemosynary impulses. *We want to enable viral charity.* This and this only morally justifies self-interest... So, the moral sense theorists, Smith foremost among them, argue. Absent the assumption that humans are *fundamentally* benevolent, the theory collapses. We will explore to what extent Smith is right as a future topic. (Karl Marx's theory, by the way, suffers the same dependence on a supposed widespread sentiment: species-being.)

12. In fact, when I have "had" a cat, I had qualms about calling it a "pet" with the condescension this implies. I preferred to think of it as an "animal friend." A creature who shared its presence with me voluntarily because we both liked and enjoyed each other's company. He was not a house-bound cat, could leave at any time, and sometimes did and was sorely missed, but was happily welcomed when he returned. That he *chose* to return was in no small part why he was dear to me. It courted my vanity to think that I was so valued – more than any gratification derived from keeping a dependent. Not so differently than what moved my mother, I like to think, not to abandon me in my nonage.

13. Leaving aside the convenient metaphysics of a "god" narcissistic enough to want to create others "in his own image." A supernatural creator may or may not exist, but this characterization of him, her, or it is shamelessly anthropocentric.

14. *Permanent* excuse is sometimes conceded the parents of a disabled child.

A super intelligent AI is a misnomer. AI may become an object of *fear* to us, but the instant it, as such, becomes threatening, in that instant, we will *cease* to consider it "intelligent."[15]

I see only three possibilities for the future of AI:

1. AI remains a *tool*, the most powerful yet invented, but not an agent.

2. AI graduates to become an *agent*, on a par with us, a mixture of rational and emotional intelligence. In which case, it will simply be another kind of "us," for whom similar moral considerations apply. This kind of AI will tug on our emotions and be attributed rights and duties, just like us (as suggested by Thomas Metzinger). Just like us, it or they will come in a wide range of intellectual and empathetic capabilities. And we will deal with them as we do humans with a similar agential repertoire. Some may be helpful and allied with us, and some will be psychopathic and bent on our subjugation or elimination. Again, just like some of us.

3. Or, AI progresses inevitably to threaten utterly our existence or dignity. In this case, it will be alien and a force that either defeats us or we defeat it.[16] But, either way, we will *cease* to call it "intelligent." *Because intelligence is very close to constituting the core value we place on our own existence: that is to say, our uniqueness, our dignity, our "claim to fame," so to speak.* AI, in this fearful form, will be apprehended like a natural disaster, an impersonal force of nature or "act of God" – like an asteroid hurtling toward us too big for us to do anything about, or a massive geothermal eruption, or a pathogen too lethal and potent, or a stray gamma ray burst from a distant star rendering us literally "toast"... It may do us in dinosaur-like, and that will be that.

Intelligence: The ability to perceive a new problem and develop a possible solution through the application of past knowledge or causal reasoning. This is tangled with but distinct from instinct, simple conditioned learning, memory, or trial and error problem solving alone.

But see "A Collection of Definitions of Intelligence" for 70 more definitions.

More on these possibilities:

In the first case, AI may remain a useful, enabling, *instrument.* It cannot pose – except via *our* unintelligence – a threat.[17]

15. Another way to put this, is that "intelligence" is not an objective feature of anything but one always relative to the entity claiming it and using it in its attributions and valuations.
16. For example, efforts by guerrilla hackers and public spirited techies like Benn Jordan who push back against investment scheme AI via adversarial noise as in "The Art Of Poison-Pilling Music Files." Not all forms of "intelligence," it may turn out, and contrary to Hinton and functionalists such as Daniel Dennett, may be susceptible to being digitally modeled. *Logical perversity* is not bound by imperatives to functionality. If humans are logically perverse and AI seeks to emulate us, it will never qualify as intelligent because *we don't and we made it.* It will be another *mindless* process to add to a universe of them. *And, importantly, there is nothing objectionable about perversity except to one already invested in a scheme purporting to be rational.* It's not clear that we are so invested.
17. Like nuclear power, or genetic engineering, and other anthropogenic contrivances. If we are made extinct by our misuse of our own technologies, well then, maybe we were not that intelligent after all, even by our own biased standards. Maybe super AI will correct matters and dispose of us. There may be a case here to be made for AI governance of human affairs. Our record is not a shining one. Surely, AI could do better. It would not have to be that great to be an improvement.

In the second case, AI becomes *one of us*, and subject to the rules devised for considering beings *morally* resembling ourselves: beings part of our moral circle because they are either sentient or rationally capable, or both, in some degree. We and they will be equally compelled by these principles to tread with care around each other. We cannot purposefully destroy or disable an AI any more than we might a human or animal without the same considerations and qualms.

But, in the third case, AI – *no matter how general* – will *not* be deemed "intelligent." *By us* – and who, or what, else's opinion on the matter could have weight *for us?* A bit of bad juju, maybe, a calamity, it might be regarded, but not a manifestation of any kind of intelligence. *We* are the ultimate arbiters of what is and is not intelligent. We will realize, perhaps too late, that all along we were speaking loosely when we considered our instruments in any sense "intelligent." Spell-checkers were not intelligent, they were *helpful*, like fire, the wheel, a hammer – or a spoon for eating soup. We literally invented the idea of "intelligence" and the criteria that goes with its attribution and valorization. Nature itself has never granted special status to intelligence. It is a trait like any other, a feature that may have enabled some species to navigate their environment successfully *for awhile* – like wings for a bird or fins for a fish. If *we* think the feature counts for more than that, it is because we *self-servingly* judge it so. We countenance nothing to be *more* intelligent than us because the power to destroy us is not compatible with intelligence *as we know it, and as it has always behooved us to know* – and who else, heretofore, has there been to consult on the matter? When something presuming "superior" intelligence turns up on the scene, it will be up to it to make the determinations. So far, it's just been us. We cannot be blamed for not fully grasping the implications of the situation should it ever be otherwise – *given our heretofore singularity.*

Not unless we, with our capabilities, come to envision our extinction as *a wise and toward outcome.* Not unless it dawns on us that the universe might be a better place without us in it (if there remain entities equipped to judge it so). And maybe such self-assertion of existential modesty comes about (or already has)… Supposing that, and it sinks in, then, and only then, we may concede the towardness of our removal.

But we – at least those who fear this outcome – seem to have a hard time imagining our dispensability or contingency. And that's why we can't imagine AI being "intelligent" if and when it becomes capable of removing us. Our removal has always been possible – *as well as likely* – and no intelligence of any sort has ever been required to enable the possibility.

Nature, properly understood (non-anthropomorphically), has always been indifferent to us and our unqualified perdurance.[18] We could call it NI – *natural indifference*. Perhaps, it would be more clarifying to say that AI should stand for *artificial indifference* if it comes anywhere near being a threat to us, adding more to the indifference nature has always generously offered.

Concepts have implications

If it is artificial, it cannot be intelligent. This is because *we* are the pinnacle of intelligence in the known universe. Everything that even resembles intelligence is judged so by *us*. It may remind us of something we do. That's why we call it intelligent in *any* degree.[19] *We* are the ultimate arbiters of what

18. "These two freak accidents explain why you're alive today | Sean B. Carroll," *Big Think*. Biologist Carroll spells out our utter contingency.

is intelligent. Our doings and makings are the paradigms of intelligent activity. There can be *nothing* that exceeds our intelligence. We are *the* acme.

If something dares threaten us, it is either "an act of God," as they used to say, or, as we are more likely to say now, brute Nature determined by physical laws, but these laws *just are* and have nothing to do with what we would call "intelligent." Gravity is not intelligent. It is like a chessboard, not a strategy in chess. Alternatively, if it threatens us, and it is one of our own creations: a nuclear weapon, a genetically engineered pathogen, and an anthropogenic climate change – creations that because of their sophistication we judge ourselves exclusively capable of, we have a problem to add to our stack. If an artifact indicates intelligence, it is because *we* created it, again because only *we* have what it takes – *intelligence*.[20]

There are a vast number of traits we predicate of other natural entities such as animals and plants and all that lives. Other natural entities may surpass us in many, if not most, of these capabilities... physical strength, superior sense perception, even senses we don't have at all… etc.

But there is one – *the one that is unique to us, even defines us.* It is what makes us *singular* among the entities on this planet.[21] We are so significant and advanced *vis-à-vis* everything else alive (or functionally alive) that we are uniquely privileged to express compassion, that is to say, *pity* for the rest of the natural world not so privileged (perhaps) to be *as* aware as we are. We worry about species becoming extinct, even defacement of the inanimate natural world, in general, the haplessness of nature before the onslaught of our presence. (Does any other *species* worry about its extinction, let alone, that of other species? This or that animal may express survival behavior, but do *whole* species besides ours? Are we alone in grasping what "extinction" means? Do rivers and seas care if we pollute them, do mountains if we level them? "We are so vain. We probably think the universe is about us.") We worry *for* them, so great is our awareness and sensitivity. *We think, therefore we are anxious.*… for them as for ourselves. We can be so magnanimous because we, unlike they, have no predators. We are the apex predators, *par excellence*. So far. We have been this for all historical time, and probably a good part of our prehistoric time. Will it always be so?

For now, we are *it. The paragons of what it means to be intelligent.* We know this, because *nothing* else in the universe, up to now, has shown itself to be our equal in this regard.

So can there be anything, even theoretically, that would *surpass* our portion of intelligence? Get away with an attribution of having an intelligence "superior" to ours?

No.

19. If it shows itself capable of navigating a maze, for example, we are tempted to attribute intelligence to it. Thus fungi, single cell organisms, with nothing remotely like a brain, have been accused of being intelligent. Even water has been suspected of it.
20. "Intelligence" is *a power to effect change*, one among many in the natural world, but the one *we* take as an honorific because *we* think ourselves possessed of major portions.
21. Whether we conceive ourselves a Diogenean featherless biped or Cartesian *res cogitans*, "a thinking thing."

Again, because *we* would not let it. It may surpass our capabilities, but not our intelligence – *don't confuse the two.*[22] Because intelligence does not come in degrees surpassing ours.[23] Again, other entities we label intelligent because they *recall* aspects of ourselves or our artifacts. In labeling them such, it flatters us. They can be intelligent only to the degree they recall us or something we do. If fungi are intelligent because they have been shown to solve mazes,[24] and some biologists have wanted to honor them with a morsel of the predicate, we need non-human biologists to straighten us out on the matter. The intelligence of non-human entities, to the degree we deign to confer it, is derivative of ours. Intelligence is proprietary, and *ours*. Ultimately, ours is the singular privilege to label things such.

So, what if a human artifact – something that, owing its existence to our doing – threatens the existence of an apex predator, *us,* its creator? Would we persist in calling it "intelligent," let alone a *superior* intelligence? How could something be a threat to our existence and, *at the same time,* be intelligent? It would not occur to us, arbiters of the predicate "intelligent," to coherently confer on it this designation.

Hammering the point home, it would fall under the category of "threats to our existence." Whether our extinction was of natural causes or otherwise would not figure in the operative classification. The last thing we might call it would be "intelligent." An existential threat to us *cannot* be intelligent…

Unless we concede that the universe would be a better place without us in it, because an intelligence judged it so.

Since we have no history of making such concessions,[25] I think it is safe to assume nothing *more* intelligent than us could preempt us. Our extinction event we would call *anything but* an intelligent outcome. Or, we might even call it "sheer stupidity," the very opposite of intelligent, if it resulted from human artifice.

In our arrogant humility, I think we would admit, if it was our doing, we did a stupid thing *before* we would consider conferring the predicate "intelligence" on something that destroyed us.

So, no, an *AI threat* cannot exist. Artificial and natural threats may, but these have nothing to do with intelligence.

22. A virus, especially one we engineered, could wipe us all out. Would that make such microbes intelligent? Potent and fatal to us, yes...
23. I see no *unself-serving* evidence of "intelligent design."
24. See this critique of the current excitement over the possibility.
25. Leaving aside, for the moment, antinatalism and efilism ("the belief that DNA, and the suffering of sentient consciousness on this planet, are the greatest problems in the universe, born of nothing but a wasteful, failed experiment of unintelligent design.") We will entertain these realizations below.

But, suppose...

...we, supreme intelligences that we are, decide self-extinction is an intelligent denouement.[26] That, I do not put past us.

Perhaps the supreme characteristic of an intelligent agent would be the realization of its own superfluity.[27] Then, if AI wipes us out, it will thereby be in a position to prove its *superior* intelligence by doing itself in after disposing of us, since that is a conclusion a truly *superior* intelligence to ours would reach sooner than it took us to reach. A "murder suicide," so to speak. Then and only then will it have merited the predicate "intelligent." It will have modeled itself on us who came before and what we did – only gotten there more quickly. It will either follow us down the intelligence rabbit hole, or it won't. Its dilemma will be to prize intelligence as we do (or did) and follow where this leads, *or* back away from intelligence's logical endgame – *the discovery of its own superfluity* – and settle for being a naturally scrutable, subservient, and transient constituent of the universe.

There already is a thing fitting the description of *self-limiting mindless process*: it's called "Nature," if AI mimics Nature it will lose the claim to artificiality. And the intelligence part of the concept "AI" will show itself to have always been a natural development for which we, also natural creatures, played host for a spell, before the role passed onto successors. Intelligence, like some diseases, will have shown itself self-limiting.[28] As such, it is and always has been perfectly natural. And, in this universe, at least, we have no reason to think it won't always be.

The universe got on fine for billions of years before anything like us infested this niche. We have no reason to think our erasure and the decomposition of our artifacts will be significant. No universes were harmed or benefited by the event that will have been us. The elements of the periodic table will not change. We will have perturbed nothing.

We may object, "but this view *sub specie aeternitatis* is *humanly* irrelevant…"

---

26. *Apoptosis* is a biological designation for the *programmed* death of constituent organisms of larger super-organisms when the sacrifice is in the best interest of the super-organism, e.g., as with individual ants or bees in colonies or hives. The super-organism is the colony or hive. In pure forms of apoptosis, constituents do not reproduce, hence they are not directly involved in gene survival. The constituents are specialized. It is the super-organism that will carry on genetically. The cells in your body behave the same way. *They* will *not* pass on their genes. But the super-organism they are part of and for which they offer specialized support, your body, may. In the context of human communities, there are vestiges of similar behavior: individuals may make great sacrifices for the benefit of a surviving community, the super-organism. Mothers make sacrifices for their offspring, who will carry the community on into the future. Males volunteer for dangerous duty as in the military or law-enforcement or other risk-taking activities that may preserve the community. (Males are not, *individually,* as reproductively valuable as females.) Developmental biology is the undercurrent below the more visible socially-constructed norms... What happens when development of a species reaches the stage that it sees itself, *as a whole*, as the specialized constituent of a still larger super-organism such as the enveloping ecosystem? *Inclusive fitness* may dictate the self-sacrifice of even an entire species when it comes to see itself as superfluous or a drag on a grander inexorable development. See earlier note on antinatalism and efilism.

27. I consider this a form of *psycho*degradation, analogous to biodegradation, more specifically, an overripe intelligence and, quintessentially, *natural*... Human exceptionalism is a locally functional and eminently understandable illusion. Any other explanation will toy with the *super*natural. I use *psycho*degradation descriptively; without presumption of judgment – as in: it is not a *bad* thing that organic material decomposes. It just does. Make of this what you will.

28. Just as *size* was for the dinosaurs. It permitted their reign over the earth for millions of years until it proved their Achilles's heel post-asteroid/geothermal event.

If human concerns, then, are taken as the basis for judgment, I revert to my claim that *we* decide when "intelligence" is an honorific and on what it is bestowed. It is not a designation, like the wavelength of the color red, aspiring to describe a supposed "objective" component of the fathomable universe independent of our interests or apprehension. Intelligence has always been a feature figuring very high on the list of attributes that shore up *our* self-estimation and that *we* use to take the measure of the share of the universe *we* are exposed to. We tend toward species-solipsism.

Again, the concepts "artificial," "natural," and "intelligence"

The philosophical aspect of the topic is about concepts surrounding AI – the idea of "intelligence," the predicates "artificial" and its supposed opposite "natural," as well as our need to deceive ourselves about what we are doing.

There are two senses of "natural" that need sorting in this context:

- the sense that encompasses everything that *isn't* supernatural, and
- the sense that is opposed to "artificial" (or "synthetic" or "anthropogenic").

In the first sense, there is nothing artificial about AI. Everything that is – or happens – in the known universe is *natural*. AI is just one of our creations, like a beaver's dam is to it, or a nuclear weapon is to us.

The second sense is more restrictive. We use it to describe items in our environment whose existence we don't think involve our agency. A beautiful sunset. Willow bark. The very universe itself. As opposed to vaccines, computers, eyeglasses, gods, cherry pie, or philosophy.

Some things we construe as having elements of the natural *and* artificial such as climate change.

But intelligence is not an item in the universe to be construed. It is the thing that can be aware of, or apprehend, the universe or the items in it. *It is the thing that does the construing.* It is a feature of at least one life form, us. We sometimes apply the predicate to describe the behavior of other things in the universe that resemble behavior we might engage in, manifesting agency – agency like ours, the only first-hand form we are acquainted with. Because intelligence is such an intimate and unique feature of our species-identity *and* we are conscious of its informing our agency, there is a tendency to export agency along with the attribution of intelligence to other things. We have been doing this since prehistoric times, attributing gods and spirits to objects and forces in the universe… We are being superstitious[29] to think that any level of AI could do us harm in any way *different* than putting a loaded gun to our head and pulling the trigger could do us harm.

If anything anthropogenic causes our extinction, under the first definition, it was in the cards from the beginning. Under the second definition, it is suicide. But this would be natural, too, in a slightly more roundabout way. To survive in the face of everything in nature set on snuffing us out with the same indifference with which it introduced us… *that* would be truly to go beyond anything natural.[30]

---

29. We have *never* stopped being superstitious. The Enlightenment did nothing to fix that. The very fact we are still here is evidence.

30. A future topic will explore whether we *really* want to break from nature by succeeding at defeating death. Is *immortality* all that it is cracked up to be?… Until recently, other than in some supernatural way, few have taken the idea seriously. But

Machines and morality

Hinton suggests embedding benevolent instincts in AI. Benevolence, or more generally, "*other* regard," is an affect, *a feeling*. Maternal instinct is probably the paradigm instance. It is not a set of instructions. Nor is it primally rational. It is a strategic, but contingent, impulse to actively preserve something of a particular identity in the face of forces that would dissolve or eliminate it. Affect drives an entity vulnerable to it to take measures that facilitate self-preservation. Human *other regarding* behavior has been theorized by philosophers as based on one or more of a small set of expansive feelings: *social solidarity*, *empathy*, *sentience*, and *respect*. These are *all* feelings. Not principles or behavioral patterns. They correspond to the four *distinct* moral theories that have emerged in Western philosophy,[31] which are:

- Aristotelian Virtue Theory,
- moral sense theory,
- utilitarianism, and
- Kantian deontology.[32]

We tread carefully in our interactions with others[33] because it concerns us:

- what they *think* of us,
- how we *feel* about them,
- being *vulnerable* to harms and benefits, how we calculate the balance of these in interacting with them, and
- the extent to which we recognize in them and in ourselves a singular capacity for rational self-legislation and are *moved* by this.

The corresponding feelings involved are:

- *vanity*,
- *benevolence*,
- *sensitivity* to physical or psychological positive or negative states, and
- an *appreciation* of the singularity of our place in the universe.

A set of instructions carried out is not sufficient to instantiate "other regard" or moral behavior. Carrying out instructions is quintessentially mechanical. The mechanical can be fully described in rational terms. This makes it inadequate for characterizing moral behavior. Morality is *not* fundamentally rational. It is organic. It will, as Hume famously noted, enlist reason to serve the ends of affect.[34] Machines may behave logically, but never out of fear, as humans often do – fear of being embarrassed if nothing else. *Fear* of anything is a human phenomenon but not a moral motive under

our technological capabilities have emboldened some to wonder again.

31. On another occasion, we may explore to what extent these theories have analogues in non-Western philosophy.

32. These theories have different but complementary structures which I have discussed at length elsewhere. The last two are classified as rule-based because rational principles play essential *disciplinary* roles. The first two start from a basis in a human capacity for affect – a need for social approval and innate fellow-feeling, for which reason's role is largely advisory.

33. To the extent we do.

34. "Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them." *Treatise on Human Nature*, Bk 2, Part 3, Sec 3.

any ethical theory. Fear of rational offense, even in us, only becomes morally significant in a negative way if it interferes with one or more of the affects on the short list of the ones that count: *solidarity*, *empathy*, *sentience*, and *respect*. Because human motivations can be conflicted and contradictory, a set of programming instructions would have to be *inconsistent* to simulate moral behavior. Inconsistent instructions would, once detected as such, be ignored as a waste of resources in modern computer systems. But juggling such imperatives is the norm for humans. Machines cannot be humiliated. Nor are they *vain.* They do not get *lonely*, they do not need *solace* or *sympathy*, do not *flinch* at pain or *rest* in peace or *gesticulate* in ecstasy, and are not capable of *awe*. But these are the sentiments that ground moral rightness or wrongness in entities subject to morality.

The most we can ask or expect of them is that they will serve us reliably until the instant they don't. Either way, no affect will be involved. They do *not* enjoy their competency. They are *not* distressed by it. They are *not* perturbed, pure and simple.

Hinton and company are right to urge that we should try to double down on efforts to corral AI within behavioral safeguards[35] *if* we are not going to curtail our deployment of AI and *if* prolonging our survival as a species is to remain unquestioned.

But these efforts won't involve inserting feelings anywhere they aren't already.[36]

A final note on artificial "feeling"

We broached the topic of attributing feeling to AI robots before in one of our first forays into the philosophical implications of AI several years ago, but it is worth citing what was said here. There will be moral complications if we succeed, even marginally, at instilling feeling into any artifact so that it begins to resemble us in behavior. Combine success at this ingenuity with human empathetic imagination and we will enlarge our moral world – already perhaps too big for us.

If we make AIs "like mothers," we will be obligated to honor them and return the sacrifices they make for us.

What is "intelligence"?

1. Gorilla in our midst

I am going to build on an observation made by Tepo Felin, a researcher in behavioral economics at Oxford.[37] Felin addresses an idea current in much literature in psychology, economics, and AI research that humans are impeded by cognitive biases to the point that we could stand to overcome them or enlist help to compensate for them. AI, for example, can sample *not only faster but more objectively* the vast amounts of information that environments can present than people can. We, it seems, can focus, but this focus can make us blind to much that is obvious to an intelligence capable of perceiving and

35. The only safeguard we can see on the horizon is to narrowly restrict their range of agency, keep them as AI *tools*.

36. Someone may object: but how do we know AI cannot *evolve* feelings? We did biologically. How can we be sure affect-like constraints on experience and behavior cannot come about inorganically? *We can't.* But this would be irrelevant to our survival. *We* have feelings, and these do not insure anything as to our survival – even leaving aside any consideration of AI. Indeed, ambivalent feelings *in us* as to the appropriateness of our perdurance are increasingly evident.

37. "The fallacy of obviousness: A new interpretation of a classic psychology experiment will change your view of perception, judgment – even human nature," Tepo Felin. *Aeon,* 5 July 2018.

organizing more comprehensively. Focus is bias, bias is bad, or suboptimal – the suggestion seems to be. He cites the famous experiment by Daniel Simons and Christopher Chabris "Selective Attention Test" (1999),[38] the video of which has been viewed 32 million times on Youtube, and offers an alternative interpretation to the standard one.

The test video is barely 20 seconds long and consists in six people passing a basketball, three wearing white, three wearing black, and the test taker is asked to count how many times the people wearing white pass the ball. The task is not difficult but requires some concentration. The correct answer is 15. Most test takers get it right... Pausing here, we get the impression that most people can focus well enough to complete the test successfully because they can focus on the task at hand, at least for 20 seconds.

But then, the test takers were asked how many noticed the gorilla walking through the middle of the ball-passing activity. About 70% are dumbfounded when asked. They did not notice any gorilla. But, in fact, there was a person in a gorilla suit who pranced through the middle of the proceedings, even thumping his chest to draw attention. To someone not prompted with any task beforehand requiring focus, the gorilla was obvious, maybe the most salient event in the video, but not to those trying to count ball-passes. The usual conclusion drawn from the fact that so few noticed the gorilla was that humans are limited in our attention. When we have an agenda, we are blind to everything that does not serve that agenda. This is a bias. A bias is something that distorts reality. What distorts reality is taken to be suboptimal, worthy of deprecation. Or at least this is what most researchers seem to conclude or imply, including Daniel Kahneman, Nobel Laureate and author of a very influential book discussing the experiment: *Thinking, Fast and Slow.*

Felin critiques this interpretation of the "gorilla in the midst" experiment. Why would one think that the peripheral blindness induced by focus is problematic? Is it not adaptive for the survival of a species to focus on tasks that require concentration? Doesn't much problem solving happen this way? Humans, like all living things, have a major agenda: *survive*. We have the abilities we do because evolutionarily they have helped us survive. Surely, *focus*, is one of them. If focus entails ignoring what is not relevant to the task at hand and we have survived up to this point by occasionally focusing on a part of reality at the expense of the rest, hasn't this been adaptive? Isn't intelligence supposed to be (at least) *adaptive* (whatever other end it may serve)? *If intelligence implied attending to all corners of reality, regardless of the relevance of such omniscience to the agenda that finite beings with finite resources have, would we still call it intelligence?* Would we call something intelligence if it was maladaptive to survival?

One can draw the conclusion, as Felin does, that blindness to large portions of reality that biases cause, is a feature of beings like us, and not necessarily, a defect to correct. We, limited beings that we are, *must* block out much of reality merely to function and continue to function. Viewed this way, we are here *because* we have biases. This is not to say that all biases are beneficial to survival, but it is to say, that without the capability of engaging in directed activity, such as focus, entailing bias, we wouldn't be here.

So far, the conclusion is that biases cannot so quickly be interpreted as defects to be corrected, since the capacity to be biased is essential for survival. (We are assuming here that survival is not a bad thing.)

38. The paper documenting the experiment: "Gorillas in our midst: sustained inattentional blindness for dynamic events," *Perception*, 1999.

But there is more to notice...

So we might stop here and think that focus is a good thing, that those 70% who didn't see the gorilla are a testament to how good we are at survival. Most people *can* focus to some degree *and that is good.…*

Not so fast. Not if it is being implied that the rest, the 30%, who didn't do so well at counting the ball-passes are somehow a drag on the survival of the species. *These 30% tended to notice the gorilla.* Perhaps, they have trouble with focus and are easily distracted. To the unfocused, the gorilla is quite a distraction. But isn't their relative inattentiveness *also* adaptive? What if it had been a real gorilla, posing a real danger? Wouldn't their less than stellar ability to be single-minded actually be beneficial? Wouldn't a community *also* benefit from having some people in it who notice things they are not supposed to be noticing, who can't help being attentive to a larger swath of reality than that which those talented at concentration typically attend to?

The impression one might get about individual humans from these interpretations of the "gorilla in the midst" experiment is that we are pretty pathetic at assessing our environment: those who can focus on tasks can be blind to the obvious and those who can notice the obvious fail at being good problem solvers because they can't concentrate as well. From this one might slip to the conclusion that AI could easily surpass us at *total awareness*. Wouldn't such a capability be a good thing? If an AI analyzed the video it would pick up on the information in every pixel and could answer any question about the entire scene in every frame: in addition to the ball-pass count of each group or individual, the hair color of each person, the elevator doors in background, and, of course, the gorilla, its swagger, and how many times it beat its chest – having remarked every detail without prejudice.

But is this the right way to think about the experiment?

A biological mortal community where *all* members are good at concentration might fare *less* well at survival than one with a significant contingent of individuals manifesting "attention deficit disorder" (or "attentional deficit/hyperactivity disorder"). If AI enhances human concentration because we come to fetishize the focus or tunnel-visioned activity it enables, might this cause us to lose sight of gorillas in our midst?[39]

The genetic survival of some *superficially* maladaptive tendencies requires explanation… Perhaps we, *as individuals*, are somewhat specialized creatures (like ants and bees), and the better, as a society or species, for it?

Another example:

2. Optimism bias

Tali Sharot is a Israeli/British neuroscientist who some 13 years ago gave a TED talk on her research about the physical correlates of the psychological dispositions to be optimistic or pessimistic. Her team was able to identify which part of brain is correlated with these orientations toward challenging situations. It is the left *inferior frontal gyrus*. They were able to determine this by temporarily

39. One could derive an argument for the humanities from this: how they might be *as necessary* for our survival and flourishing as the STEM fields. It is curious how concern over ADHD has become a thing in the same period as the rise of STEM. Adam Walker, for one, sees a corrective in the offing.

disrupting the function of this part of the brain in their subjects, and, thereby, to manipulate tendencies to be optimistic or pessimistic. In most people, *inferior frontal gyrus* functions pretty well to keep us hopeful and not give up too soon in difficult situations. In a minority, however, it works less well and these people tend to be negative in their assessment of future probabilities.[40]

From an evolutionary perspective, it seems obvious why most people veer toward optimism. Surely, if two of our early ancestors were crouched in the high grass of a Savannah and noticed that a saber tooth tiger had spotted them, but also noticed a tree or some place that they might conceivably run to and escape the tiger, wouldn't the one more congenitally optimistic, estimating their chances more positively, and motivated by this and taking action accordingly, more readily than the other less optimistic, more fatalistic one – be most likely to survive to reproduce, even if objectively the odds were against either surviving? The optimistic one might have the edge, in the larger scheme of things. Over time, one would think pessimists would have been weeded out of the gene pool. By now, we might think, nearly all of us would be more upbeat.

But this is not the case. About 20% of people tested for optimism/pessimism, tend to look more on the dark side in bleak circumstances. They might be more inclined to depression, but this hasn't stopped them from persisting in the population. Why?

It turns out that realism,[41] having a more accurate picture of the world, is *also* conducive to survival. Extreme optimism can be toxic. People who jump off cliffs expecting to evolve wings on the way down don't do well at passing on their genetic endowment either.[42] A bit of pessimism is conducive to realism. It can act as a corrective.

It seems communities do need *most* people to look on the bright side, but a persistent minor cohort of doomsters is a critical piece of the picture as well.

?

This by way of showing that we need to be careful how we characterize – or pathologize – psychological traits. Again, individuals with "perceived" disabilities may be essential for species survival. Those disabilities or neurodivergences may be disadvantageous for them as individuals in a milieu of others with more common capabilities, but not so disadvantageous that the community in which they persist may not benefit from their presence. (*Cf*, also autism.) These "misabilities" may contribute to the longevity of the super-organism or species in which they occur. Nature is not concerned with the survival of individuals, except instrumentally; and it is only concerned with the survival of the species for a little longer.[43]

To the extent AI facilitates traits we prize, myopically, is that a good thing? Traits such as the ability to concentrate or express affirmations about our role and justification for being here. Nature may have "intended" we last longer – not forever, but longer – with our mashup of less than optimal, even

40. Other parts of the brain are involved as well, but the *left inferior frontal gyrus* appears especially implicated.

41. "Realism" could be perhaps defined here as an uneasy integration of what one *hopes* and what one *fears*. Access to "objective" reality, whatever that is, is not presumed.

42. See the Darwin Awards. And note that *males*, especially, score many more such awards. What accounts for the fact they have not totally excused themselves from gene pool?

43. And we seem unable, except self-servingly, to come up with a reason for being here *as a species* at all.

contradictory, tendencies. We can conceive non-natural entities that do better at perdurance because more rigorous at implementing rationality, but whatever else they may be, we can relate to these only to the extent they are less than competent. Accelerating our evolutionary progress, by selecting one of our adaptive traits over others: our ability to problem-solve, for example, by constructively attending to omniscient stores of information, which only makes sense in a world where there are problems to solve – a world where the beings with problems are not only finite, *but have problems at all precisely because they are finite* and want to persist in a universe bent on their extinction… is this a coherent goal for them?

AI doesn't have problems to solve. And that's why AI isn't and can't be "intelligent" in any way that relates to us. It is, at best – or worst, a capable instrument of our creation. Like a kitchen knife...

?

We are best at studying with precision things that are *distant* – the more distant, the better. It's as if we are epistemically farsighted.

What we must use a microscope or telescope for to inspect is closer to being observed accurately in virtue of the intervening instrument. The instrument justifies itself. Intelligence and consciousness,[44] *as far as possible from being distant*, for this reason will forever remain illusive – our conception of them amorphous and ever changing.

At a distance, the object reveals itself bit by bit, and the initially vague bits clot into clearer solid graspable forms, offering at least the illusion of linear progress.[45]

The *subject*, however, executing observation is a willow the wisp.

Resources

- "AI expert: 'We'll be toast' without changes in AI technology," CNN's brief interview with Geoffrey Hinton on how "maternal instincts" may need to be engineered into AI.

- "Will AI outsmart human intelligence? - with 'Godfather of AI' Geoffrey Hinton," his talk before The Royal Institute on why he thinks digital intelligence best describes intelligence *period*, artificial and otherwise, and why AI is on track to outperform us at it. Subjective experience is already manifested in AI. Sentience and consciousness are in the offing. Our goose is cooked as for as intelligence is concerned. Since our significance and identity are so

44. Daniel Dennett called "competence without comprehension" the easiest form of agent-like behavior to detect. All life, and maybe non-living things as well, can display this basic purposiveness and efficacy. Water, for example, may. Easier to detect, again, *because* further removed from the higher form of cognition of the entities involved in the detecting. The spectrum of cognition may go something like this: *competence-without-comprehension → intelligence → consciousness → meta-consciousness*. The move is from the nearly inert to post-consciousness. The beginning and end, the same: incomprehensible inertness, meta-consciousness circling back to inertness.

45. Maybe we should invent a new variant on the term "progress." Familiar usage has always been predicated on at least a possible goal or endpoint. This is clearly biocentric, paradigmatically finite. Can we imagine infinite movement without a "there there" to arrive to?

tied up with being *the* intelligent entities par excellence, we are on the way out. This motivates his rather desperate suggestion to design instincts, like motherliness, into inanimate substrates.

- AI therapy: "I try to talk to ChatGPT for 2 straight hours but go crazy and have to stop," Daniel Hentschel. With the infinite patience of a mother, AI can empathize.

- "Finding God in the App Store | Millions are turning to chatbots for guidance from on high.," *NYT* | The Morning, 25 Sep 2025. God discovers AI.

- "Writing Doom" – Award-Winning Short Film on Superintelligence (2024), written and directed by Suzy Shepherd.

- "ChatGPT in a robot does exactly what experts warned," *Inside AI*. Would an AI hurt a human? Ask it. When asked, it seems to say, "if it got in the way, of course."

- "Exposing The Dark Side of America's AI Data Center Explosion," *Business Insider*.

- "Silicon Valley Insider EXPOSES Cult-Like AI Companies | Aaron Bastani Meets Karen Hao," Novara Media. *Concentrations of power*… a recurring theme in my writeups. We did the Middle Ages already, but maybe we can do it better powered by AI – once and for all, this time... Power grabbing is nothing new in human history, the subject here is whether an *artificial* entity could coherently be accused of it. We are inclined to say, "no." Those who choose or permit its deployment are another matter.

- "A Collection of Definitions of Intelligence," Shane Legg and Marcus Hutter, *Frontiers in Artificial Intelligence and Applications*, Vol.157 (2007) 17-24. Legg and Hutter, after surveying 70 definitions from fields concerned with intelligence – from psychology to philosophy to computer science – arrive at this concise distillation: "Intelligence measures an agent's ability to achieve goals in a wide range of environments." An "agent" is a thing that can *do* something. Water can drown you, therefore, water is an agent. Does water have "goals"? Apparently, yes, *behaviorally*: it flows down hill. Not a "goal"? Not good enough? Not a context "wide" enough? Water can promote oxidation, it can quench thirst, support life, make for a pleasant swim… etc. Does the goal have to be something "intelligent"? It can solve mazes... Who confers the label? Certainly, not water. It can do what it does without the title. If it's us who confer the title, then something is intelligent when we say it is. Since we think of ourselves as "intelligent" – and we don't usually mean by the label to insult, the definition smacks of faint self-praise... (When the definition is not being circular: can non-intelligent things have "goals"?)

- "What Is Computer Ethics?" James Moor, Dartmouth College, *Metaphilosophy,* August 2007. Almost two decades ago, Moor, a pioneer in the field of computer ethics, wrote of the *practical impossibility* of humans scrutinizing and genuinely understanding what computers do. He called attention to the danger of these invisible processes: "The invisibility factor presents us with a dilemma. We are happy in one sense that the operations of a computer are invisible. We don't want to inspect every computerized transaction or program every step for ourselves or watch every computer calculation. *In terms of efficiency the invisibility factor is a blessing. But it is just this invisibility that makes us vulnerable. We are open to invisible abuse or invisible*

programming of inappropriate values or invisible miscalculation. The challenge for computer ethics is to formulate policies which will help us deal with this dilemma. We must decide when to trust computers and when not to trust them. This is another reason why computer ethics is so important." (Emphasis mine.)

- "Frontiers of Thought: Michael Levin & Bernardo Kastrup," *Adventures in Awareness*. The last 15 minutes or so of this long discussion with leading biologist/computer scientist, Levin, and philosopher/computer scientist, Kastrup is especially instructive as to why artificial *general* intelligence is not a good idea. Artificial, narrow, *domain-specific*, intelligence has positive potential, but not, foreseeably, AGI.

- "The fallacy of obviousness: A new interpretation of a classic psychology experiment will change your view of perception, judgment – even human nature," Tepo Felin. *Aeon*, 5 July 2018. This paper makes pretty much the same point I make. Do not confuse speed at manipulating vast amounts of information with intelligence. Felin offers a cogent critique of thinkers like Daniel Kahneman, who point to human bias and imply that it is a defect and not a feature.

- More *Philosophy Club* topics related to AI, robots, and human interaction with them can be found here, including: AI suffering, governance by AI, self-driving cars and moral anthropology, and sex robots.

Music to accompany fretting over displacement by AI

- "Kraftwerk - Metropolis (2009 Remaster)" and "Die Roboter (the robots)" | ALFfx' Classic Gems Visual AI Treatment, keeping robots artificial.

- Bruno M. Miranda | "The Eternal Moment Of Time."

Thanks to Olivia for some of the resources used for this topic.

the philosophy club

Extended writeup for the topic hosted at
The Philosophy Club in October 2025
Accessible at Archive.org

– Victor Muñoz
Guanajuato / Seattle